AFRL-IF-RS-TR-2007-26
Final Technical Report
January 2007

NEW METRICS FOR CHARACTERIZING AND PREDICTING NETWORK BEHAVIOR

University of South Carolina Research Foundation

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. T018/00

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Rome Research Site Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-IF-RS-TR-2007-26 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/s/

JAMES L. SIDORAN
Work Unit Manager

/s/

WARREN H. DEBANY, Jr.
Technical Advisor, Information Grid Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JAN 2007	Final	Jul 04 – Dec 06

4. TITLE AND SUBTITLE

NEW METRICS FOR CHARACTERIZING AND PREDICTING NETWORK BEHAVIOR

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA8750-04-2-0260

5c. PROGRAM ELEMENT NUMBER
62301E

6. AUTHOR(S)

Joseph E. Johnson, Vladimir Gudkov, Chin-Tser Huang, Cilia Farkas and Duncan Buell

5d. PROJECT NUMBER
T018

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of South Carolina Research Foundation
901 Sumter St., Ste 511
Columbia SC 29208-0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Advanced Research Projects Agency
3701 North Fairfax Dr.
Arlington VA 22203-1714

AFRL/IFGB
525 Brooks Rd
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2007-26

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA# 07-028

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Networks are systems of point (nodes) with connections among some pairs of nodes measuring the degree of linkage. Networks represent an entire problem domain of many of the most difficult and unsolved mathematical problems. The objective of this effort was to formulate a foundational structure for networks and specifically develop new mathematical metrics for the description of networks in order to usefully describe both the static and dynamic properties of networks. Specifically, these new metrics provide a means of monitoring networks such as internet traffic over time by identifying anomalies, malicious processes, and abnormal network behavior. The criteria used for establishing network metrics were: a) well-defined mathematically, b) lossless in the description of a network, c) hierarchical in providing a sequence of numerical metrics of decreasing importance, d) intuitive in order to guide the use of the mathematical network expansions and associated metric values, e) descriptive of the inherent topology of the network and strengths of connectivity, f) sufficiently fast computationally in order to be dynamically useful as a tool, and g) ideally distinguishing types of metrics that are: 1) network invariants, 2) variables which have dynamic behavior, and 3) variables which are chaotic or random. Results include: a) finding network metrics that satisfy these criteria, b) building the computer software to derive such metrics for general networks, and c) testing limited internet traffic with this software.

15. SUBJECT TERMS
Network metrics, network topological structures, entropy measures, Markov processes, predicting behavior, connectivity matrix

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			JAMES L. SIDORAN
U	U	U	UL	32	19b. TELEPHONE NUMBER (*Include area code*)

Table of Contents

1 THE PROBLEMS, OUR OBJECTIVES, AND IMPACTS 1

1.1 Background and Problem Description 1

1.2 Objectives and Goals 5

1.3 Expected Impact: 6

2 TECHNICAL APPROACH: PROCEDURES & METHODOLOGY 7

2.1 Detailed Description of the Technical Approach 7

2.2 Software Written, Tested, and Reviewed 11

3.0 TECHNICAL MATHEMATICAL BACKGROUND 14

3.1 Connections of Markov Transformations to Lie Groups and Monoids 14

3.2 Connections of Markov Monoids to Networks 14

3.3 Comparison with Current Technology 16

3.4 Explicit Computational Procedure for Proposed Network Metrics: 17

3.5 Network Dynamic Tracking 18

4 RESULTS, DELIVERABLES, & FINAL PRODUCTS 19

4.1 Deliverables Description 19

4.2 Technology Transition and Technology Transfer Targets. 19

4.3 Practical and Computational Considerations 19

4.4 Interpretation and Discussion: 21

4.5 Results of Monitoring Internet Traffic 23

5 CONCLUSIONS 25

5.1 Resulting Network Metrics 25

5.2 Further Analysis Needed 25

5.3 Difficulties Encountered with Real Time Deployment 25

6 REFERENCES 26

Abstract

Networks are systems of points (nodes) with connections among some pairs of nodes measuring the degree of linkage. Examples include communications networks (the internet and phone networks), transportation networks (airlines and roadways), utility networks (electrical, natural gas, water-sewer grids), financial networks, social linkages, and disease transmission networks. Nodes could be airports and the linkage as the number of passengers flown between pairs of nodes. Networks represent an entire problem domain of many of the most difficult and unsolved mathematical problems. We seek to formulate a foundational structure for networks and specifically develop new mathematical metrics for the description of networks in order to usefully describe both the static and dynamic properties of networks including identification of changes such as system failures, reorganization, and malicious processes. These network metrics are to satisfy a set of criteria and specifically to link the established mathematical foundations to practical applications. An example of what we seek was achieved with the Fourier expansion of sound waves into component cosine waves. Specifically, these new metrics are to provide a means of monitoring networks such as internet traffic over time by identifying anomalies, malicious processes, and abnormal network behavior. Our criteria for the network metrics are that they shall be : (a) well defined mathematically, (b) lossless in the description of a network, (c) hierarchical in providing a sequence of numerical metrics of decreasing importance, (d) intuitive in order to guide the use of the mathematical network expansions and associated metric values, (e) descriptive of the inherent topology of the network and strengths of connectivity, (f) sufficiently fast computationally in order to be dynamically useful as a tool, and (g) ideally, distinguishing types of metrics that are (i) network invariants, (ii) variables which have predictable dynamic behavior, and (iii) variables which are chaotic or random.
We have been successful in (a) finding network metrics that satisfy these criteria, (b) building the computer software to derive such metrics for general networks, and (c) testing limited internet traffic with this software. Our work is founded upon our discoveries: (1) that every possible network is in one to one correspondence to a specific Markov transformation. This provides a solid mathematical foundation and connects general network theory both to continuous groups and to Markov processes. (2) This led to our ability to define general entropy (including Shannon & Renyi' information) functions on the rows and columns of the associated Markov transformation and thus to distil the topological order/disorder representing the flows into and out of a node into entropy values. (3) Our next advance was that the ambiguous numbering of the nodes allows one to sort these entropy values into a histogram representing the entropy spectra of the network thus obtaining two curves (incoming & outgoing) representing the entropy spectra of the network. (4) Finally we studied the profiles of different networks and their changes over time by monitoring these entropy spectra to determine what is normal for a given network and thus to identify anomalous behavior and exactly what nodes are involved. To accomplish this we developed a full software package in both Mathematica and JAVA for users to deploy with networks of their own interest and expertise. The final products of our work include (i) technical and white papers, (ii) patent applications, and (iii) operational network monitoring metrics software at www.exasphere.com.

1. The Problems, Our Objectives, and Impacts

1.1. Background and Problem Description

1.1.1. Foundational Laws of Nature are Based Upon Vectors: All of the foundational laws for our physical world are formulated in terms of vectors which give the positions and velocities of particles. All of classical mechanics rests upon describing the state of any physical system as a vector of such metric values and then predicting the time evolution of this vector. [1] Even with the advent of the theory of special relativity and quantum mechanics, the formulations are still in terms of vectors (although in an infinite dimensional space). [2] A vector is an ordered set of numbers such as (2, 4.6, -3….). Even extremely complex systems are well represented by the linear superposition of vector forces between component objects that give the behavior of any component as responding to the sum of all imposed forces. Our inanimate physical world has been very successfully described by these vector based laws to an accuracy that is unimaginable.

1.1.2. Networks are not Vectoral but Matrix Based: Yet with living things and in particular social structures and activities, a new thing has emerged, the network. Networks were first studied extensively over the last three centuries by mathematicians addressing problems such as the traveling salesman problem (who must visit each of a set of cities once and only once in a minimal path). [17] Over the last half century social scientists have studied the networks formed by people into groups, cliques, and clusters including criminal and gang activity as well as organizational and social structures. Then with the advent of highway traffic and telephony over the last century and the internet over the last decade, a domain of problems has emerged that is of a very different nature and complexity than that encountered in traditional science. This new domain is characterized by relationships and connectivity as the fundamental characteristic, not as a vector but as a matrix or square array of values as we will now show.

1.1.3. Description of a Network as a Connection Matrix C_{ij}: Fortunately there is a mathematical description that is clear, unambiguous, exact, and easy to understand. If we number the nodes or points of the network with the numbers 1, 2, …. N then we can represent a network as a matrix or array of values between nodes i and j as C_{ij} = strength of relationship from i to j. C is called the 'connection' or 'adjancy' matrix. C_{ij} might represent the number of airline flights per week between city i and city j or the total number of passengers flown between the two nodes in a day. Or C_{ij} it might represent the amount of money transferred in a given month from bank account (or accounting category) i to account j. And, central to our considerations, C_{ij} might represent the number of emails or bits of information sent from computer i to computer j (without regard for the stops in between the two). The critical point is that ANY network can be described by such a matrix $C_{ij.}$ This matrix is a square (two dimensional)

array of numbers as opposed to a vector which is represented by a one dimensional sequence of numbers.

1.1.4. Essential Properties of the C_{ij} Network Matrix: The C matrix is not just any set of numbers but must satisfy the following: (1) Zero Diagonal: As it is normally not meaningful to talk of the relationship of a thing to itself, the diagonal values of C are not defined and are usually taken as zero (C_{ii} =0) For example it does not make sense to speak of the airline flights from a city to itself. (2) Non-negative Values: The values of C are not negative but are zero or positive numbers (C_{ij} >= 0). That is because we consider the minimal relationship to be none at all. By the last analogy, one cannot transport a negative number of passengers from one city to another. (3) Large: A major characteristic is that networks are very large and thus C_{ij} is a huge set of values (N^2- N for N nodes where we subtract the N diagonal values). It is easy to envision a computer network of a million or more users with a resulting C matrix consisting of a trillion values. It is this astronomical size that overwhelms even any imaginable supercomputer. (4) Time Dependent: Many of the networks are ever changing and thus we must consider them as time dependent ($C_{ij} = C_{ij}(t)$). Consequently the overwhelming number of values within C for one time is compounded again astronomically as we consider that we have a succession of values at each time where the number of airline passengers changes each day and the internet transmissions change every second. (5) Node Numbering (C) Ambiguity: In order to just write the C_{ij} matrix, we must assign numbers to the nodes. Yet if another person were to number the nodes, one would end up with a C matrix with the values in very different places. In fact for large matrices, it is generally impossible with any known computer to even compare to see if two networks are the same. In general there are an astronomical number of different C matrices (N * (N-1) * (N-2)…*1 = N!) that describe the same exact network. This ambiguity arises from the fact that there is generally no normal order or sequence for the nodes that is not arbitrary in some sense. Consequently although C_{ij} describes a network exactly and uniquely, there are a vast number (N!) of C_{ij} that will describe that same network so one has an essential ambiguity due to the lack of a natural way to number the nodes.

1.1.5. Operational Issues in Defining C: (6) Connection Value (Weight) Definition: In establishing the weight values in C, it is important to consider what will be the exact definition of the network, specifically, how to define the numeric values of C_{ij} in terms of real world parameters. Consider banking transactions where one could say that a transfer from i to j simply gives a C_{ij} = 1 and otherwise a zero. But with a little thought, one realizes that a lot more information is contained in the C matrix if the values indicate the total amount of the transfer. But is $1,000 really 100 times more important than $10? When we refer to the income of a person we often say he or she makes a 'six figure' income. An equivalent way of saying this is to take the logarithm of the value of transfer which can be shown to be indicative of the power of 10 that will give the value in question. Thus one

might wish to use the logarithms of the component transactions in order to construct the values of C_{ij}. Thus there is no predetermined method of assigning weights. It is the domain expert who must decide what exact function of the network weight data will best represent the topological connections for the use and analysis intended. (7) Time Window for Summation: In general one finds that a network is usually formed from a set of transfers that occur over time. For example with the number of airline flights or passengers, with financial transactions, or with internet data transfers, one begins with a database file of records which have the form: (a) date-time, (b) node i, (c) node j, (d) weight of connection, (and possibly (e) a data transfer type indicator). Then one is to sum the weights over some period of time, δ, from $t+\delta/2$ to $t+\delta/2$. If one chooses too short a time period δ then the resulting C_{ij} matrix is almost all zeros and one gets meaningless results. If one chooses too long a window of time then all temporal changes are washed out and averaged leaving no dynamical structure. An exact analogy exists with photography where too short an exposure (such as a trillionth of a second) will result in an image with only a few dots of light while too long an exposure (such as several weeks) results in an overexposed blur consisting of averages of all motions. Thus it is a complex problem to decide what time interval will best capture representative information at a given point in time. Again this must be decided, as in the weight case above, by an expert in the given domain of knowledge. (8) Data Collection: Finally, all the above points are moot if the requisite data cannot be collected. In many cases this is easily achieved as with banking transactions or airline flights as such data is routinely collected in electronic form. Privacy issues can be managed by obfuscating the identity of the nodes in a renumbering of the bank accounts or airlines and storing the 'true identity' of each node in a separate secure database. This is a necessary procedure anyway because the nodes must be labeled with sequential integers (1, 2, 3, …) in order to construct the C matrix. This renumbering hides the true node identity in the associated correspondence table. Indications of attack or system failure for specific nodes can be relayed to the holder of the secure database which gives the true identity of each node. However, most cases are much more complex and often it is impossible to gather all the data that one needs to define all parts of C_{ij}. One example is in capturing the packet transfers among nodes when the transfers occur in a foreign country, or in a small office group that has an internal router that does not register the transfer above the indicated level. Other serious problems can result from temporary IP addresses and other ambiguities in computer traffic that block either the identity of nodes or make the determination of the linkage unobtainable. A similar difficulty arises in financial networks when some transfers are made 'under the table' using unregistered diamonds or transfers through countries and banks that do not report the transfer.

1.1.6. Comparable Scientific Problems: Problems similar to some of these core problems have arisen before in science. The most obvious is when

one has a gas of N identical molecules. Not only can one not order the molecules to distinguish one molecule from another, one has such a vast size (10^{24}) that one could never take the measurements of where each molecule is located nor could one even write down all of the equations, much less solve them! But a more critical point is that if one could do this, the result would be meaningless of a prediction of where each molecule is at some future time. Instead one needs a new and totally different set of 'metrics' for the gas that give its overall holistic state and dynamical development. Such a problem is solved with the variables of thermodynamics and statistical mechanics such as temperature, internal energy, volume, pressure, entropy, etc. that summarize the uncountable number of individual particle properties and which ignore which particle is doing exactly what at a given instant of time. [3] These 'thermodynamic metrics' provide a holistic view of the gas and its behavior. Thus here we see the core of the problem: We must find (a few) new holistic variables (network metrics) that summarize the essential nature of the network. These new metrics must distil down the essence of the network connectivity (topology) and disregard the vast number of unimportant data values.

Yet networks normally have no concept of distance (an email or financial transfer is just as close to a person in China as it is to a person down the street). Without distance one cannot define a 'volume' or a 'pressure' (force per unit area). Also there is no conserved quantity such as energy and thus there is no natural definition of 'heat' or 'temperature'. In fact 'equilibrium' is not even well defined in general on networks. This leaves us with 'entropy' (or its negative information) which can be defined as the disorder (or order) in a probability distribution. However, the network in itself does not have a natural content of probability distributions, only arbitrary normalizations.

1.1.7. Importance of the Problem: We see that a vast spectrum of modern problems center on understanding the status and dynamical behavior of networks. Our entire national and world economic system of communication and information linkages, financial transfers, shipments of goods, transportation of people, delivery of utilities, and the contagion of disease can only be managed and understood if we are able to understand the behavior of networks. All of these social-economic-communication-transportation problems are network based and not vectoral in nature. One of the most daunting of these problems is the emergence of the internet for computer and personal communication including the remote control of devices by internet, by software or persons from a distant site often without our knowledge or understanding. Foremost among these problems is the emergence of computer bugs, worms, viruses, attacks, and an entire spectrum of malicious processes requiring something akin to the biological defenses necessary for the maintenance of life forms [6,8,9]. One example of a network that can wreak havoc on modern society is a network of persons who have criminal or terrorist objectives. But if we

cannot maintain secure communication and control for our military and corporate complexes then we become subject to attacks and destruction of our social order to an extent never before conceivable. The problem of understanding, monitoring, tracking, and securing networks is of the greatest possible importance to the security of our nation, world order, and the very survival of any form of advanced civilization [16,18].

1.1.8. Conclusion: Essentially all aspects of our social fabric of civilization are composed of networks. The problem of understanding networks is thus of the greatest possible importance for the future stability of all complex social systems. The foundational description in terms of the C matrix is well defined and unique (within a few adjustable parameters and certain difficulties of data collection). Privacy issues can be managed at least technically although we must deal with such issues politically. So what is the problem? Although the problem resides in the all of the eight (8) issues discussed above, it primarily is rooted in the shear volume of network data and the fact that any one value is of the same importance as every other value.

1.2. Objectives and Goals

1.2.1. The Central Goal – Network Metrics: The central goal of our work is to identify functions (which we call network metrics) of the values in the connection matrix, C_{ij}, that summarize the data into useful variables as described in the requirements above. Central to such identification is that the variables are sensitive to overall structure of C without being sensitive to the details of the data, as with thermodynamic variables. Critical to our goal is that the network metric variables that are to replace C must have a firm mathematical foundation and yet be 'useful' in the practical sense. These network metrics with their hierarchical order of importance are to replace the astronomical number of equally important values in the C matrix.

1.2.2. Examples of Similar Metrics in Science and Mathematics: This is a difficult and somewhat ambiguous requirement but we certainly understand it as we have diverse expansions in physics, chemistry, and engineering such as the expansion in Fourier series for sound (or other waves), multipole expansions for charge and mass distributions, and a large number of expansions in orthogonal functions. The central characterization of such expansions is that one begins with nearly infinite data and one obtains a few numbers that are in ordered importance. With the Fourier expansion one begins with millions of changing amplitudes of the audibly distinguishable frequencies and reduces this first to a fundamental tone or note of frequency f_0. Then one determines the amount of sound at twice that frequency '$2f_0$', and sequentially in multiples or overtones of the basic frequency nf_0. Other notes and deviations arise as much higher terms with very small coefficients. The same is true with the multipole expansion for a charge distribution where the first term is the total charge; the second is the dipole moment, then the quadrapole moment, etc. One of the first solutions that occur is one

which has been studied for decades: that of determining the eigenvalues of the connection matrix. It is known that the eigenvalue spectrum is not unique for each unique topology but the main difficulty is that such a computation is extremely laborious for large matrices thereby reducing the usefulness of this method. Furthermore the C matrix is still ambiguous as the diagonal values are undetermined thus the eigenvalues and eigenvectors are totally ambiguous and arbitrary making the whole procedure questionable.

1.3. Expected Impact

1.3.1. The implications and impact of a set of characteristic metrics that satisfy all of the above criteria would be truly phenomenal and far reaching. In particular these metrics would be of even greater importance if they were divided into (a) those that are essentially invariant, (b) those that have discernable dynamical behavior, and (c) those that are chaotic or random. This would allow one to probably ignore those that are random or chaotic and to concentrate on the deviation of the invariant metrics or the alteration over time of those that had simple time evolution properties.

1.3.2. There would be mathematical implications as well as if we could develop a set of metrics with a firm foundation in the description of the underlying topology – and that would be our aim. This is because the essence of the network is the essence of the topological connectivity and to have functions that measure the order and disorder of the organization of connections would be the most desirable end results.

1.3.3. In the last analysis, even if all of the tasks above are satisfied and fulfilled, it will take a vast effort to examine different types of networks to ‘profile’ the new metrics in terms of ‘normal’ and abnormal behavior and even to the extent of measuring ‘how abnormal a given topology is from the norm, in terms of deviation of the metrics from their normal profile.

2. Technical Approach: Procedures & Methodology

2.1. Detailed Description of the Technical Approach

2.1.1. Initial Technical Approach: The analogy with thermodynamics reduced the problem to the utilization of entropy as probably the most useful holistic function of the C matrix components. The discussion above left other thermodynamic variables without a foundation in this domain (lacking a distance metric and lacking a conserved quantity like energy as well as a lack of a natural 'equilibrium' state for networks in general). Our further discussions and investigations strengthened our premise that entropy is exactly the correct variable in that it measures the order or disorder in a probability distribution. It is precisely such an order/disorder metric that we would like to have that reflects the structure of the underlying topology of the network. But the problem was that C had an undefined diagonal and furthermore C did not contain probability distributions. We worked for over a year testing mutual entropy functions and studying the results of arbitrarily normalizing the C matrix columns or rows to probabilities (which was achievable as the values are all non-negative). Thus our early work with one of the component teams (with Dr. Gudkov) studied this approach [12,13] along with cluster and clique analysis to identify certain topologies. With this team we also utilized complexity theory and made two important discoveries: (a) We experimentally found evidence that the internet, in the sense of complexity theory, had a dimensionality of about 10 to 12 dimensions or independent parameters (out of about one hundred variables in the IP routing structure that could be possible). This implied that other degrees of freedom would be random or chaotic and thus unpredictable. (b) The second discovery was that mathematically, Hausdorf dimensionality was directly related to the order of the Renyi' entropy. The combination of these results indicates that there are about 10 or so degrees of freedom and thus that Shannon and Renyi' entropy up to about order 12 should give useful information about the internet if we could figure out how to form such functions from the C matrix. Next we realized the importance of the results by Kolmogorov that all information about the network structure could be represented in the information (or entropy) functions, in the forms of Shannon and Renyi', of the C matrix if the C matrix could be converted to probability distributions. Thus although we had strong indications that entropy functions of the C matrix would be the optimal network metric functions, we did not know the precise form as we did not have probability distributions from the C matrix. Related work and investigations by another component of our team (Dr. Buell, Dr. Huang, and Dr. Fracas) centered more on practical applications of the identification of attacks and anomalies using just the number of transmissions into or out of a given server (that is just the total sum of elements in the C rows and C columns). These processes ignored all higher order connectivity and order/disorder structure.

We emphasize that to the very lowest order of approximation, one can just look at those total connection values which give the total traffic into or out of each node. Many research programs simply work with those values over time as this alone is a vast quantity of data. The main thrust in such research is to identify the anomalous patterns in that data alone over time.

2.1.2. Technical Breakthroughs & Primary Discoveries: After extensive meetings with consultants and experts and after reviewing the literature, our work centered on how to extract the essential topological and structural features of the network using the one well defined representation of any network, that is the C matrix. We first determined that no such 'summary metrics' were known. A lot of work had been done by taking the eigenvalues of a C matrix with either zeros or ones on the diagonal. But this method was so time consuming computationally that it was useless for large scale networks. It was also known that these 'eigenvalue metrics' were not sufficient to distinguish some networks which were topologically different. Recalling that the C matrix is a square matrix of non-negative values and with no definition for the diagonal, we sought ways that we could utilize entropy functions to abstract the network structure (topology & connectivity) with probability distributions built from the C matrix in order to find a few functions (network metrics) of the C matrix that captured the 'essential structure' of the network and be insensitive to the massive data of what every individual node was doing. We reasoned furthermore that the C matrix could be thought of as a measurement of flows that the network represented at that instant such as people in transit on planes or electricity in transit between stations.

2.1.3. New Insight: However, without some kind of creative leap forward, there did not appear to be any obvious solution for the desired network metric functions. The breakthrough came as a result of prior research on the subject of continuous groups of transformation and the Markov transformations. It was realized that this C matrix was, (apart from the its diagonal), identical to the generator of an infinitesimal Markov transformation [7,4]. Since this Markov generator matrix had non-negative off-diagonal terms and diagonal terms that were the sum of the other elements in each column, one only had to alter the C matrix to have the same diagonals. Then each altered C matrix had fixed diagonals and became the generator of a continuous Markov transformation. Consequently we obtained our primary result: that every network (that is every C matrix of non-negative values with an undetermined diagonal) is in one to one correspondence with an infinitesimal Markov transformation and conversely [5]. Such an altered C matrix is now exactly determined and the associated Markov matrix is of the form $M(a) = e^{aC} = 1 + aC + a^2C^2/2! + \dots$. The importance of this is two-fold: (a) on the one hand we now have connected three distinct domains of mathematics – network theory, continuous groups of transformations, and Markov theory. (b) But from the practical point of view, the Markov transformations, M, have the remarkable feature that each column constitutes a probability distribution

that can be used to support an entropy function. (c) Not only that but this M transformation has a formal interpretation of the flow of a conserved probability among the nodes at the rates given by the off-diagonal values of the C matrix. Thus (d) it follows that the columns of M^c are probability distributions that reflect exactly the topological flows into a node indicated by that column.

2.1.4. Row Transformations Equally Important: Also, another type of Markov generator emerges if the diagonals of C are determined to be the negative of the sums of the row values (rather than column values as before), then the probabilities now given by the rows of the associated M^r matrix give the flows out of (rather then into) the associated node for that row. For example one could formulate [10,11] the second order Renyi' entropy for column j as $E^c_j = \log_2 (N \Sigma_i M^c_{ij}{}^2)$. The Shannon entropy for column j would be defined as $S^c_j = \Sigma_i M^c_{ij} \log_2 M^c_{ij}$. Likewise each Shannon and Renyi' entropy could be computed for the M^r matrix. Consequently, we have determined how to form the entropy functions in a unique manner and furthermore this method gives entropy functions that are exact distillations of the topology of the network for flow out from and flows into each node. This method reduces the C matrix of N^2- N values down to 2N values which is a very substantial compression of information. Still for a network of a thousand nodes we are left with two thousand values rather than 1 million but how do we begin to follow even this set of summary metrics?

2.1.5. Meaning of the Entropy Metrics of the Markov Matrix: To understand what these entropy metrics mean, we must first realize that the above procedure takes the altered C matrix and generates an associated Markov matrix (one where the columns are probabilities and one where the rows are probabilities). The probabilities in the matrix M^c can be shown to be the relative flows into each node from that node representing the column in question. Thus a given column of M^c represents the relative transfers into that column, in proportion to the connections given in the generating connection matrix. Likewise the row values of M^r can be show to represent the relative transfers out of the node associated with that row. Taken together, we see that all of the transfers, as generated by the C connection weights, into and out of each node, are measured by the probability distributions of the columns and rows of M. Since these transfers are exactly representative of the exact topology of the network, it follows that these probabilities exactly measure the topology.

2.1.6. Definition of Entropy on Rows & Columns of M: Now knowing that these probability vectors represent the topology into and out of each node, one can ask what does the entropy of such a probability measure. The entropy function acts to measure the disorder in the probability distribution. So if a vector of probability has all the same values (e.g. (0.2, 0.2, 0.2, 0.2, 0.2)) then the distribution is uniform and the disorder is maximum, just like having dirt evenly distributed in a room. But if the probability is

concentrated in one place (e.g. (0,1,0,0,0)) then the disorder is a minimum just as though all the dirt were swept into one pile and put in the trash can there. Thus the entropy functions measure the extent to which the probability is distributed or concentrated – both for incoming and outgoing transfers. Thus the entropy is an accurate summary of the topological structure of M and thus of C.

2.1.7. Further Data Reduction: Still one is left with 2N different entropy values for the incoming and outgoing entropy metric for each node. For a network of a million connections this gives two thousand values that are constantly changing and consequently one is still overwhelmed with data in spite of the massive reduction in the number of values that have to be tracked. The solution to this lay in the realization that we do not really care which exact node does what as long as about the same number of nodes are doing the same inputs and outputs. For example with telephone networks, it does not matter if Bob calls 10 people from 8am to 9pm and then gets off the phone and Jack calls 10 people in the next hour. In either case the phone 'topology' is unchanged as it does not matter which node connects to 10 other nodes. Based upon this we realized that it is only the pattern of the entropies that mattered and we were able to reformulate this as how to how to represent the general structure of all of the entropies in the following manner.

2.1.8. Entropy Spectral Distributions: The next discovery thus centered on the problem with how to represent the entropy spectra. We realized that since it did not matter which node had what entropy, then we could sort the entropies in order and form a spectral curve using a non-increasing function. The shape of this curve then indicates the topological structure of the network flowing into the nodes. By using the same sort order, we can obtain another spectral curve representing the row entropies of flows out of the nodes. A look up table (database) is maintained to allow one to go back to find which node corresponds to each point in the graph. Thus one now has two entropy spectral curves that encapsulate the topological structure of the network. Then if two nodes exchange roles in having the same outgoing or incoming entropy, then the spectral curve will remain the same. The nodal orders can chance over time but one is only interested in the shape of the curve and to what extent it deviates from the normal curve. The next problem is then to know what is normal and what is abnormal for a given type of network?

2.1.9. Experimentally Determine the Normal Network Spectral Curve: The exact topology for each type of network is not something we can predict at this point. Rather, one simply experimentally determines the shape of the E^c and E^r entropy spectral curves over time by measuring the network C matrix sequentially over time and computing the associated entropy curves again and again. One can then determine the average shape of the entropy curve along with one standard deviation. Then at a future instant, one can overlay

the E^c curve at a given instant of time on top of the average E^c curve to see if there are any points which differ by more than an accepted variance. At those points which so vary, it means that the underlying network topology differs from the normal topology. Using the lookup table, one can then identify the nodes for which there is a variance and then investigate whether there is a true attack or system failure.

2.1.10. Normalization of the C Matrix – Volume of Flows: We next realized that if all transfers in the matrix were twice as great, then the topology would be the same and the spectral curves would be the same also. However, the parameter 'a' in the expansion of the M matrix in terms of the C matrix would have a different value for these two cases thus making the two spectral curves more difficult to compare. One can see that the parameter 'a' in the expansion, tells us basically how many connection levels we wish to incorporate in the transfers since each power of the C matrix in the expansion of M will represent the connections to the connections to the connections etc. Although we are primarily interested in the topology of the relative flows, it is true that if a network matrix C suddenly has all of its values double or triple in size, then this represents an important event. This is similar to having the same investment pattern for stocks and bonds yet increasing the level of investment in all exactly proportionally. In the stock market this is referred to as the 'volume of trading'. Thus we need a way to normalize the C matrix to a standard value and to record and plot this normalization as the "Amplitude of Transactions". Using the fact that the trace of a matrix (sum of all diagonal elements) is invariant (and here the trace is also the negative of the sum of all the off diagonal elements), we choose this trace as a measure of the amplitude and thus renormalize every C matrix by dividing by the negative of the trace. This will lead to a C matrix which will always have a trace of unity. This separate value of amplitude can be additionally tracked over time in addition to the entropy metrics.

2.2. Software Written, Tested, and Reviewed:

2.2.1. Creation of Prototype Operational Code: Extensive effort was next made in rendering the mathematical and algorithmic procedures to a more powerful computer software environment. First our team spent approximately nine months in writing the code in Fortran and testing network data from the first site with whom we were able to contract (Allen University). We were able to perform internet traffic data capture and to render the algorithm correctly in order to identify attacks and malicious processes on their university network. We considered this to be a preliminary study for two reasons: (a) we were able to write the code rapidly in Fortran but did not consider this to be the final language of choice for commercial development. (b) The site data would only be available for about a 1 year period and the data was more limited than we desired for our final testing and operation. Yet this operation was successful and showed the full feasibility of the approach using rapid development techniques. During this

analysis, our techniques uncovered a number of infected computers that were being used by hackers. We also determined an even larger number of computers that were being used for illegal audio and video distribution in addition to other activities. These processes had not been uncovered by the multiple off-the-shelf intrusion software products being currently used.

2.2.2. Creation of Final Operational Code: We next programmed the algorithm in two languages: Mathematica and in JAVA with JAVA tool sets. The Mathematica version allowed users with advanced scientific backgrounds (but more limited programming ability) to do very complex analysis and visual plots of the data and for general network research. The JAVA system was designed professionally and is envisioned to be the code and tool set that is to be deployed commercially for applications to networks in all forms. Both of these systems were developed and tested using real data against each other to make sure that the algorithms gave the same answers. We have made them compatible and complementary and we are now confident in the quality of both software environments as final products and deliverables from this project.

2.2.3. Testing and Evaluation of Real Internet Network Data with the Entropy Metrics Software: Over the last nine months of the project, we set up a second server at another university (Coastal Carolina University) with a more extensive agreement for data capture and were able to capture extensive streams of network data. Although the data was still streamed after capture (and not analyzed in real time), we were able to identify abnormalities associated with malicious processes and illegal use of servers and computers on the system. The report on these results is attached along with the associated entropy spectral curves and anomalies so indicated.

2.2.4. Summary of Consultants & Experts Utilized: By working continuously with selected consultants and advisors and using reasonable scientific approaches we were able (1) to construct the list of essential and desired requirements and issues for the network metrics using analogies with thermodynamics, acoustics, quantum mechanics and other areas of physics, mathematics, and engineering (as listed above). (2) We also extensively studied the nature of the network specification problem and studied what had been done previously in the mathematics and computer science literature characterizing networks by the C matrix, and the progress that had been made in studying the eigenvalues and eigenvectors of C. (3) We held two conferences to bring together experts in intrusion analysis to understand the current state of development and technology in the field. (4) We extensively utilized six world experts for consultation on our methodologies and existing internet traffic and anomaly monitoring. (5) We presented our ongoing research both nationally and internationally at meetings and technical seminars. In each case this afforded us the opportunity to meet for several days with additional experts in the field both for criticism of our methods and for other methodologies. (6) We have published our ongoing

works in the literature and later in a refereed journal. (7) We have made our methodology known with white papers, technical papers, power point presentations, and posters on multiple levels of technicality. This material is all available on our web site. (8) We have applied for a provisional patent one year ago and now, with our final new results, we have applied for a full patent (Dec 2006). All of these methods and consultations have given us the background and external review of our methods and the formulation of our requirements.

3. Technical Mathematical Background

3.1. Markov Lie Groups and Monoids

3.1.1. We had previously shown [7] that the transformations in the general linear group in n dimensions, that are continuously connected to the identity, can be decomposed into two Lie groups: (1) an n(n-1) dimensional 'Markov type' Lie group that is defined by preserving the sum of the components of a vector, and (2) the n dimensional Abelian Lie group, A(n), of scaling transformations of the coordinates. To construct the Markov type Lie group, consider the k,l matrix element of a matrix L^{ij} as a basis for n x n matrices, with off-diagonal elements, defined as $L^{ij}_{kl} = \delta^i_k \delta^j_l - \delta^j_k \delta^j_l$ with i =/= j. Thus the ij basis matrix has a '1' in position ij with a '-1' in position jj on the diagonal. These n(n-1) matrices form a basis for the Lie algebra of all transformations that preserve the sum of the components of vector. With this particular choice of basis, we then showed that by restricting the parameter space to non-negative real values, $\lambda^{ij} >= 0$, one obtains exactly all Markov transformations in n dimensions that were continuously connected to the identity as $M = \exp (s \lambda^{ij} L^{ij})$ where we sum over repeated indices and where s is a real parameter separated from λ^{ij} to parameterize the continuous evolution of the transformation. In other words $\lambda^{ij} L^{ij}$ consists of non-negative coefficients in a linear combination of L^{ij} matrics. This non-negativity restriction on the parameter space removed the group inverses and resulted in a continuous Markov monoid, MM(n), a group without an inverse, in n dimensions. The basis elements for the MM algebra is a complete basis for n x n matrices that are defined by their off-diagonal terms. The n dimensional Abelian scaling Lie algebra can be defined by $L^{ii}_{kl} = \delta^i_k \delta^i_l$ thus consisting of a '1' on the i,i diagonal position. When exponentiated, $A(s) = \exp (s \lambda^{ii} L^{ii})$, this simply multiplies that coordinate by e^s giving a scaling transformation. The Lie algebra that results from the sum of the Abelian and Markov Lie generators is sufficient to generate the entire general linear group that is connected to the identity.

3.2. Connecting Markov Monoids to Network Metrics

3.2.1. We can begin with the simple observation that (1) since the non-negative off diagonal elements of an n x n matrix exactly define a network (via C) and its topology with that node numbering, and (2) since a Markov monoid basis is complete in spanning all off-diagonal n x n matrices, then it follows that such networks are in one to one correspondence with the elements of the Markov monoids [7,4,5]. The Lie Markov matrix that results is exactly the C matrix where the diagonal elements are set equal to the negative of the sum of all other elements in that column. Thus each such altered connection matrix is the infinitesimal generator of a continuous Markov transformation and conversely. This observation connects networks and their topology with the Lie groups and algebras and Markov transformations in a unique way. Since the Markov generators must have the diagonal elements set to the

negative of the sum of the other elements in that column, this requirement fixes the otherwise arbitrary diagonal of the connection matrix to that value also (sometimes referred to as the Lagrangian)

3.2.2. It now follows that this diagonal setting of C generates a Markov transformation by $M= e^{\lambda C}$. One recalls that the action of a Markov matrix on a vector of probabilities (an n-dimensional set of non-negative real values whose sum is unity), will map that vector again into such a vector (non-negative values with unit sum). The next observation is that by taking λ as infinitesimal, then one can write $M = I + \lambda C$ by ignoring higher order infinitesimals. Here one sees that the value or weight of the connection matrix between two nodes, gives the M matrix element as the relative infinitesimal transition rate between those two components of the vector. Thus it follows that given a probability distribution x_i distributed over the n nodes of a network, then M gives the Markov transition (flow) rates of each probability from one node to another. Thus it follows that the connection matrix gives the infinitesimal transition rates between nodes with the weight reflecting that exact topology.

3.2.3. Specifically, if the hypothetical initial probability vector is x_i = (1,0,0,0…0) then the vector at a time dt later will be equal to the first column of the M matrix, M = I + dt C. Thus the first column of M is the probability distribution after an infinitesimal time of that part of the probability that began on node 1. Likewise for all other nodes thus giving a probability interpretation to each of the columns of M as the transfer to that node. Thus each column of M can be treated as a probability distribution associated with the topology connected to that associated node and will support an unambiguous definition of an associated entropy function that reflects the inherent disorder (or order) after a flow dt. Thus the columns of M support a meaningful definition of Shannon or Renyi entropies which in turn reflect the Markov transformation towards disorder of the topological flow to the node for that column. Thus the Renyi entropy on this column can be said to summarize the disorder of the topology of the connections to that node to that order of the expansion. It follows that the spectra of all nodes reflects in some sense the disorder of the entire network. We recall that the numbering of the nodes is arbitrary and thus we can now renumber the nodes without affecting the underlying topology. We thus sort the N entropy values of the nodal entropy in descending order which gives a spectral curve independent of nodal ordering and thus independent of the permutations on nodal numbering (except possibly for some degeneracy which we address below). That spectral curve can be summarized by the total value for the entropy of all columns (since entropy is additive and the column values are totally independent.

3.2.4. If the connection matrix is symmetric then the graph (network) is said to be undirected, but if there is some asymmetry, then the graph is at least partially directed where the flow from i to j is less or greater than the converse flow. If the connection matrix is not symmetrized then one can capture this asymmetry by resetting the diagonal values of C to be equal to

the negative of all other row values in that row. Then upon expansion of M = I + λC, the rows are automatically normalized probabilities that in turn support entropy functions for each row. These row entropy values form a spectrum which could be sorted by the same nodal values (in order) that is used to order the column values. This will result in a different spectral curve that is not necessarily in non-decreasing order for the row entropies. One also can compute the total row entropy as we have done for columns. If two columns have the same entropy then one can remove some of the numbering degeneracy by using the values of the associated row entropies by using a rank ordering as we did with column values.

3.3. Comparison with Current Technology

3.3.1. Current technology is primarily devoted to the lowest level of topological connectivity, namely the shear quantity of traffic into and out of each server on the network. As such, it does not capture ANY aspect of the topology of the network or who is connected to whom. The customary metrics with the current technology are very extensive in monitoring the time series of traffic and the statistical variations into and out of each primary node or server on a network. But this is equivalent to only looking at the total sum of off diagonal elements in any row or column of the C matrix as opposed to the order/disorder patterns in the transmissions. Thus the current technology is only operational at the lowest level of computation – namely counting and performing statistical analysis. Our algorithms are capable of drilling into any level of connectivity of the network associated with any of the nodes and furthermore to fully analyze, with the entropy functions, the associated topology of the network. Thus our system distinguishes between the patterns of many transmissions to a relatively few, from a single other computer, or from a few transmissions to a very large number of computers. The entropy function actually captures the intensity pattern of such connections. This is in contrast to the traditional counting which simply counts the outgoing or incoming packets.

3.3.2. Our network metrics have been (a) well defined mathematically as they are exactly defined as well defined functions on the rows and columns of the M matrix which in turn is exactly defined in terms of the C matrix which is exactly determined by the network. (b) The entropy description is lossless for a network if one uses the sequence of entropy spectra by removing the node of highest entropy from a network and then computing the entropy spectra for the subnetwork etc exhaustively. (c) The approach is hierarchical in providing a sequence of numerical metrics of decreasing importance especially if one computes the differences in the normal spectra and the spectra at a given instant and computes the sum of the squares of differences between the two curves. (d) The proposed entropy metrics are intuitive in that they are associated with a diffusion flow among the nodes at the rates indicted by the C matrix. This intuition can be useful in many ways to guide the mathematical network expansions and give insights into metric values. (e) The entropy functions are exactly descriptive of the inherent topology of

the network and strengths of connectivity and encapsulate the measure of the order and disorder in transmissions and receptions. (f) The algorithms for entropy are very fast computationally as compared to other standard means (such as eigenvalues) because one only has to square the values after expansion in 2, 3, or 4 orders. (g) The general entropy spectral curve will be invariant for the most part (we have found) unless there are anomalous processes occurring. Much of the random processes are eliminated in the summation to compute the entropy.

3.3.3. Comparison to Eigenvalue Computation: Very extensive past work on network analysis has been done by computing the eigenvalues of the C matrix. As was pointed out above this procedure not only takes a great deal of computational time, it is also ambiguous as one has no definition of the diagonal. However, with our method there is no remaining ambiguity as one now can diagonalized the altered C matrix (with diagonals defined as described above) and this will be equivalent to diagonalization of the M matrix (since $M(a) = e^{aC}$). Furthermore, our methodology reveals the meaning of the eigenvalues as the rates of exponential decline of the flow of a conserved dispersing fluid on the network at the rates given by the network transfers. The eigenvectors are those linear combinations of nodes that have a smooth exponential decrease (at the rate of the associated eigenvalue). This shows how our work, as a byproduct, further quantifies and explains more deeply, the existing methodologies of eigenvalue/eigenvector research.

3.4.Explicit Computational Procedure for Proposed Network Metrics:

3.4.1. Begin with the connection matrix C_{ij}. It contains all network data. Compute the Markov Matrix.

3.4.1.1.By experiment or simulation, begin with an electronic database of transactions of the form: data-time, node from, node to, relationship weight = t, i, j, w.

3.4.1.2.Determine the window of time δt for grouping the transactions into a $C_{ij}(t \pm \delta t)$.

3.4.1.3.Determine whether to use the weight w directly or some other function such as log w prior to summing into the C_{ij} value.

3.4.1.4.Compute the column diagonals $C_{jj} = - \Sigma_i C_{ij}$ to be the negative of the sum of all other elements in that column. This fixes all elements of C_{ij} uniquely.

3.4.1.5.Normalize C_{ij} by dividing all elements by the negative of the trace. Store this number as A(t) which we will refer to as the amplitude of the matrix.

3.4.1.6.Determine the order of the expansion, m, and the expansion parameter λ, for the expansion to obtain M(t) as $M^c = I + \lambda C + \lambda^2 C^2 / 2! + \ldots \lambda^m C^m / m!$ These choices are interdependent because if one wants the topology to reflect only the connections to the connections and no higher order then m =2. Thus λ should be of a size where λ^2 is not vanishingly small but where λ^3 is in fact extremely small. Also the λ must be chosen so that there are no negative elements in M.

3.4.1.7. Perform the same operation for rows as was done here for columns to obtain another M which we will call M^r.

3.4.1.8. Now one has obtained two Markov matrices M^c and M^r. that contain the incoming and outgoing topological connectivity respectively. Each M contains all information in C.

3.4.2. Compute the entropy (or other statistical) spectra of M:

3.4.2.1. Each column of M^c. and each row of M^r. is a vector whose non-negative components sum to unity and thus which can be used as a probability distribution. Each is in fact the probabilistic flow of a conserved entity, at the rate of C_{ij} connectivity, in the time λ between the associated nodes. Thus the columns (and rows) support the computation of both Shannon and generalized Renyi' entropy information (entropy) functions. Choose a function type and order that will characterize the associated topology.

3.4.2.2. We have used second order Renyi' entropy in the majority of our work as it is the lowest order Renyi' entropy. Simply take the sums of the squares of the elements in a given column, multiply by n (the number of elements in a column) and take the log of this base 2 thus obtaining E^c_i (the column entropy for each of the nodes) and E^r_i. (the row entropy for each of the nodes).

3.4.2.3. Sort the resulting E^c_i in order of value and thus obtain a histogram or spectral curve of these entropies. Likewise, use this same sort order for the $E^r_{i.}$

3.4.2.4. Track these two curves over a period of time to determine the average shape (for that time of day, day of week, weather etc) and also to determine the normal variance curves at 1 and 2 standard deviations.

3.5. Network Dynamic Tracking

3.5.1. Overlay the current spectral curves for both columns and rows over the average and its variances.

3.5.2. Aberrant nodes can be easily identified by their variance from the norm. One even knows the probability of a deviation of that magnitude. In our software, one can simply mouse click on the deviant areas of the curve and the node identities are given from the lookup table of which nodes those were prior to the sort.

3.5.3. Compute a measure (such as the sum of squares of the differences) between the current curve and the average curve and take this value as $E^c(t)$. Likewise compute $E^r(t)$ thus obtaining two values along with A(t) (the amplitude that was previously calculated). Then track these three functions for a network to identify anomalous topological network behavior.

3.5.4. This process results in a reduction of the network ($C_{ij}(t)$) to just three numbers at each instant of time: $E^c(t)$, $E^r(t)$, and A(t) corresponding respectively to the deviation of the column and row entropy spectra from its normal pattern and also giving the total volume (amplitude) of network flows. The magnitude of these three functions indicate the magnitude of the deviation of the network topology from the normal at each instant.

4. Results, Deliverables, & Final Products

4.1. Deliverables Description

4.1.1. Project deliverables are (a) this final report. (b) The software in both Mathematica and JAVA source code ready for implementation by commercial users. (c) The Web site www.exasphere.com which will serve as the central point for the primary results and computer code dissemination. (d) All of the above on DVD disk delivered to DARPA and to the Air Force Office. (e) A patent application to protect the intellectual property of the software and assonated algorithm.

4.2. Technology Transition and Technology Transfer Targets.

4.2.1. A technology transfer company, ExaSphere Inc has been established in order to manage distribution of the associated software and intellectual products from this research. The patent and intellectual property will be owned by the University of South Carolina. The marketing will be managed by ExaSphere Inc. with royalties paid to the University.

4.2.2. This work has been presented in Network and Complexity Conferences as follows: (a) Toulouse France April 2004, St. Petersburg Russia September 2005, New England Conference in Boston Mass. June 2006, and in Greece September 2006. Seminars have been given at the following sites; AFOSR Rome NY, University of Arizona, University of South Carolina and the Kiawah Conference on Network Intrusion September 2004.

4.2.3. Future presentations and papers, similar to those given above, will be given to complement the ExaSphere.com marketing of the software products and patent licensing.

4.3. Practical and Computational Considerations

4.3.1. The work here has both purely mathematical and practical aspects pertaining to applications to real networks. If one only has a single C matrix and time is not involved then the following discussion on time windows does not apply. It will then be assumed that one has a data flow with records of the form: (a) network type, (b) time, (c) node i, (d) node j, (e) weight. These might be SNORT captures of internet traffic between IP addresses, financial transitions between bank accounts, power transfers among electrical grid substations, or passengers flown between two airports. The $C(t,\delta)$ matrix is constructed by summing the weights into the appropriate cells (renumbered with integers as i, j = 1, 2, …N) during a time period δ centered about time t. It is obvious that one must have a period δ which allows a 'representative' accumulation of values for the disagrigation size N. If C is too sparse, then one must choose longer time windows or one must collapse the matrix nodes by some natural methodology such as IP sectors, or flights between states and not airports. In some cases one may wish to

combine several network types using a linear combination of the contributions determined by the 'type' parameter. In some considerations, one might wish to modify the weight of the contribution such as using the log of the financial transfer. The software we have built contains loaders with such adjustable parameters. The result of this process is a C(t) with no diagonal terms. We then put this in the form of a Lie Monoid generator by setting the diagonal terms equal to the negative of the other terms in that column (and later row). We then find it useful to normalize the entire matrix to have a fixed trace of -1 or –N as this allows better control over the subsequent expansion into the Markov matrix.

4.3.2. The expansion $M(t) = e^{\lambda C(t)}$ although mathematically guaranteed to converge, have non-negative terms, and generally be Markovian, must be executed with a small number of terms if C is large. The parameter λ gives a weighting of the higher terms in the expansion where one might choose to sum up through 'k' terms. The number of such terms is the extent to which M 'feels out' the connections to the connections etc. as weighted by the parameter λ. These two must work hand in hand since it is meaningless to have a very large λ while only expanding to the first order in C. Conversely, it is meaningless to expand to many powers, k, of C while using a nearly infinitesimal value of λ since higher orders of λ will make such higher powers of C vanish. The next consideration is that although the M matrix has only positive terms when the full expansion is executed, in practice one can choose k and λ which, due to the negative diagonals of C, can give negative terms for truncated expansions. Thus the software must have error checks to make the appropriate corrections in the expansion.

4.3.3. Now having the M(t) matrix for that instant, one computes the $E_j^c = \log_2(N(\Sigma_i M_{ij}^2))$ ie the log of the sums of squares of each column to get the entropy (information) for that column representing the transfers into that node by the Markov matrix. The spectra is computed by sorting these by value while keeping a lookup table for which node goes to which original position. A similar computation is done to compute the entropies of the rows E_j^r where the same sort order is used except for removing potential degeneracies (where the column values are the same and thus not distinguished by order). These two spectral curves, or histograms, are computed for each successive time window and overlaid graphically to compare the row and column entropy profiles over time. A critical point is to realize that it does not matter that the nodes are renumbered with each window, but rather we are interested in whether the profile of order and disorder of the underlying topology is 'about the same'. Naturally some profiles for networks change from late Sunday night to rush hours at 9AM Monday. Likewise, power grids depend upon the temperature as well as the time of day. Thus for a given time of day, day of week, and if necessary for that network, weather pattern in temperature, one must learn the profile of what is normal (i.e. profile one standard deviation) for the network under consideration and then to overlay the instantaneous network spectra on this and graphically display it. One can sum all of the row entropies into a single

value $E^r(t)$ and likewise for the columns. Then one might sum the squares of deviations from normal to obtain a single value representing the total deviation of column entropies from normal (and likewise for the rows). Our software performs these computations and displays both $E^c(t)$ and $E^r(t)$ along with the overall network 'amplitude' A(t), which is the trace of the original C matrix. This gives us three curves that we can monitor over time as well as watching the current row and column entropy spectra displayed overlaid upon the normal distribution for those circumstances. One must then be able to identify where anomalies are occurring in the network for example by clicking on the associated spectral curve anomaly area. The system then finds the node identification in the lookup table thus identifying the anomalous nodes and subnets.

4.4. Interpretation and Discussion:

4.4.1. We emphasize again that the flows that are modeled by $M(t) = e^{\lambda C}$ have nothing at all to do with the dynamical evolution of the network. These metrics are used to monitor the network state and dynamical behavior but not to predict it. Rather the evolution generated by $M(\lambda)$ is an imaginary dynamical flow that would occur if a conserved fluid (probability, money, population …) were to move among the nodes at the rates indicated by the C matrix of connected weights. Thus the value of $M(\lambda)$ is that the associated entropies can be used to summarize the order or disorder of the incoming or outgoing topological connectivity of the (static) network at one given instant of time. The philosophy here is that the entropy values will capture the most essential aspects of the structure of the column and row probability distributions, and thus the topology, to that level of expansion of the parameter λ. By expanding to higher powers of C, with larger values of λ, the entropy metrics capture increasing levels of the connections to the connections etc. Also by utilizing other Renyi' entropies, one obtains other spectra and values that measure other 'moments' of the probability distributions.

4.4.2. One can also consider alternative diagonal values of the C matrix by adding the Abelian scaling group transformation generators to the diagonal values of C. These transformations destroy the conservation of the modeled flow (such as probability) and thus the resulting transformation is no longer Markovian. These altered diagonal transformations are equivalent to adding sources and sinks of the modeled fluid at the associated nodes. It is straight forward to prove that the entropy value $E(t) = \log_2(N\langle x(t)|x(t)\rangle)$ when taken to only the third level of expansion, can, with its partial derivatives with respect to such sources and sinks at the node 'j', for different initial conditions for the flow $|x(0)\rangle$ at node 'i', formally obtain the entire C matrix thus showing that the entire topology of the network is contained in the entropy functions and its derivatives.

4.4.3. When C is diagonalized, with the values leading to the Markov transformations, or to the more general values of the diagonals of the last

paragraph, one automatically gets a diagonalization of the M matrix. The interpretation of the eigenvectors is now totally obvious as those linear combinations of nodal flows that give a single eigenvalue (with exponential decrease when the transformation is Markov) of the associated probability, for that eigenvector. This follows from the fact that all Markov eigenvalues are negative except the one value for equilibrium which has eigenvalue unity for equilibrium. That means that each of these negative eigenvalues of C reflect the decreasing exponential rates of decrease of the associated eigenvector as the system approaches equilibrium as λ approaches infinity in $M= e^{\lambda C}$. This insight allows us to see that all of the Renyi entropy values are increasing as the system approaches equilibrium, which is normally the state of all nodes having the same value of this hypothetical probability. The use here of this 'artificial flow of probability under M' provides us with more than just a method of encapsulating the topology with generalized entropy values, it also gives an intuitive model for the eigenvectors and eigenvalues for C and sheds light on the graph isomerism problem (different topologies having the same eigenvalue spectra). Of course it does not resolve any graph isomerism issue associated with degeneracy of multiple topologies for a single eigenvalue spectra without altering the C matrix by the Abelian transformations.

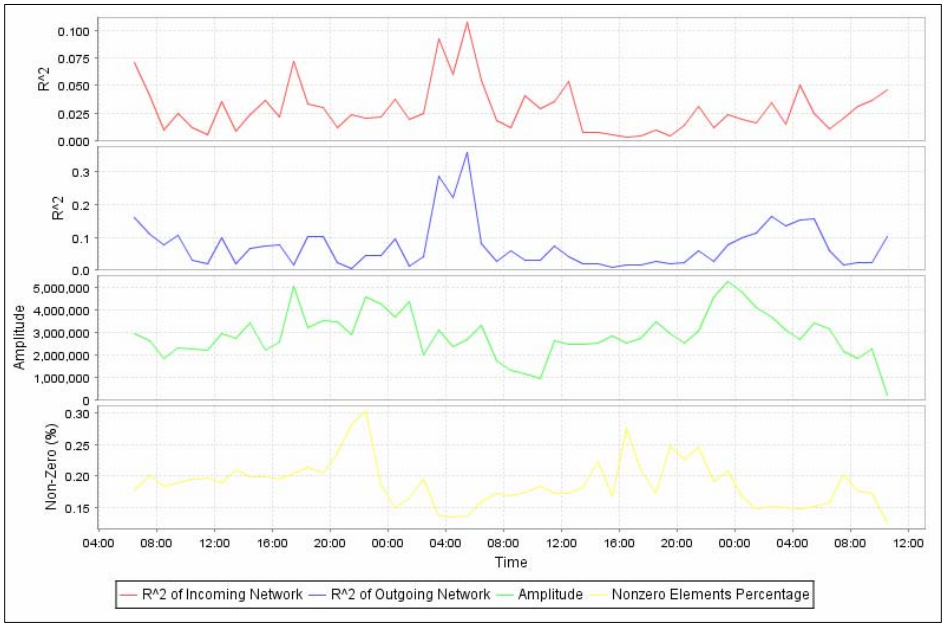
4.4.4. Based upon the arguments above, we suggest that for real networks such as the internet, that the appropriate connection matrix be formed, from source and destination information transfers, where both asymmetry and levels of connection are to be maintained in the C(t) matrix values during that window of time about that time instant. Specifically, this means that if a connection is made multiple times in that time interval, then that C element should reflect the appropriate weight of connectivity as this adds substantial value to the entropy functions. We then suggest that at each instant, the column and row entropy spectra be computed along with the total row and column entropy and that this be done for lower order Renyi entropies as well as lower order values in the expansion of the Markov parameter λ that includes higher order connectivity of the topology. We are currently performing tests to see how effective these entropy metrics are in detecting abnormal changes in topologies that could be associated with attacks, intrusions, malicious processes, and system failures. The patterns (from our simulations) of specific topologies such as rings, trees, clusters, and other structures have interesting entropy spectra. We are performing these experiments on both mathematical simulations of networks with changing topologies in known ways, and also on real network data both in raw forms and in forms simulated from raw data. The objective is to see if these metrics can be useful in the practical sense of monitoring sections of the internet and other computer networks. It is important to note that one can obtain these same metrics for subnetworks of the original network. The subnetwork would be chosen as that portion of the topology that has incoming or outgoing entropy changes that are anomalous [20]. Thus this

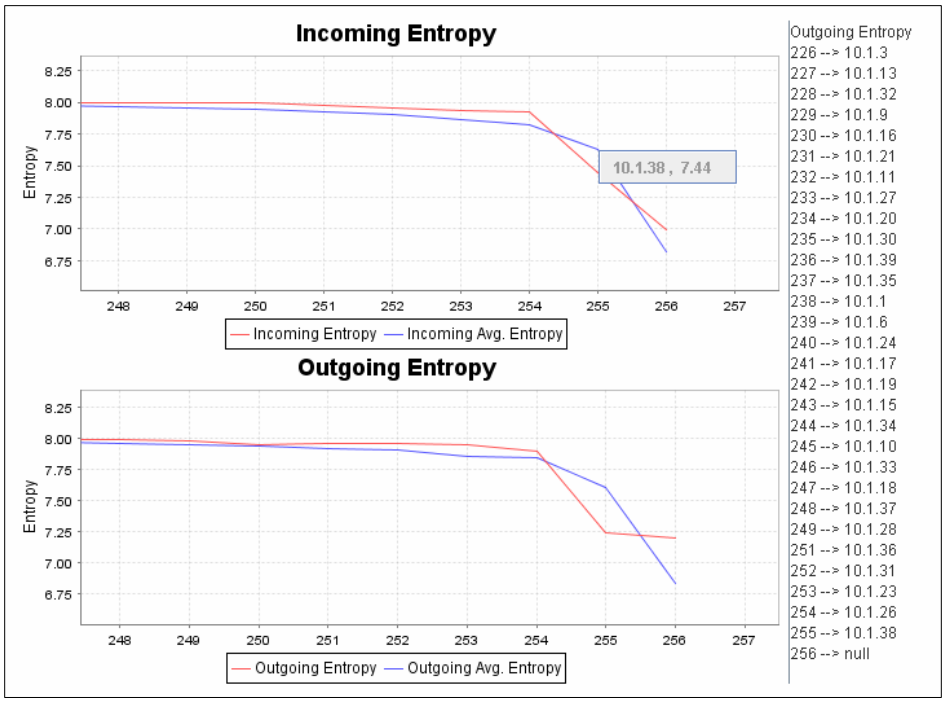
technique allows an automated reduction or hierarchical expansion methodology to drill into the network to monitor those subnets that are most dynamically aberrant.

4.5. Results of Monitoring Internet Traffic

4.5.1. The mathematical and computational techniques defined above along with the associated Markov entropy network metrics can be used to analyze the static and track the dynamic behavior of any type of network structure. This includes electrical grids, natural gas pipelines, communications networks, financial transactions, disease networks and social networks. But the network tracking that we have performed to date concentrated totally on internet traffic as defined by Snort data capture at servers of information that is sent from one IP address to another IP address. Our objective is to identify anomalies, and abnormal behavior relative to normal traffic patterns by monitoring the total column (incoming traffic) and row (outgoing traffic) second order Renyi' entropy along with the traffic volume which is independent of the traffic topology. This is similar to separating the buying pattern of financial investments from the volume of transactions on the market as two separate indicators.

4.5.2. The associated graph shows the total incoming and outgoing entropy as a function of time for a server at a university of 30,000 students and faculty. The major anomalies were identified at certain times and these were expanded to see the full entropy spectra at those times over the network thus identifying the specific nodes that had aberrant behavior. It was determined that these particular anomalies in entropy occurred for nodes that at certain times were used to upload and download large volumes of audio and video files.


R^2
0.100
0.075
0.050
0.025
0.000
R^2
0.3
0.2
0.1
0.0
Amplitude
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
Non-Zero (%)
0.30
0.25
0.20
0.15
04:00
08:00
12:00
16:00
20:00
00:00
04:00
08:00
12:00
16:00
20:00
00:00
04:00
08:00
12:00
Time
R^2 of Incoming Network
R^2 of Outgoing Network
Amplitude
Nonzero Elements Percentage


Incoming Entropy
Entropy
8.25
8.00
7.75
7.50
7.25
7.00
6.75
248
249
250
251
252
253
254
255
256
257
10.1.38 , 7.44
Incoming Entropy
Incoming Avg. Entropy
Outgoing Entropy
Entropy
8.25
8.00
7.75
7.50
7.25
7.00
6.75
248
249
250
251
252
253
254
255
256
257
Outgoing Entropy
Outgoing Avg. Entropy
Outgoing Entropy
226 --> 10.1.3
227 --> 10.1.13
228 --> 10.1.32
229 --> 10.1.9
230 --> 10.1.16
231 --> 10.1.21
232 --> 10.1.11
233 --> 10.1.27
234 --> 10.1.20
235 --> 10.1.30
236 --> 10.1.39
237 --> 10.1.35
238 --> 10.1.1
239 --> 10.1.6
240 --> 10.1.24
241 --> 10.1.17
242 --> 10.1.19
243 --> 10.1.15
244 --> 10.1.34
245 --> 10.1.10
246 --> 10.1.33
247 --> 10.1.18
248 --> 10.1.37
249 --> 10.1.28
251 --> 10.1.36
252 --> 10.1.31
253 --> 10.1.23
254 --> 10.1.26
255 --> 10.1.38
256 --> null

5. Conclusions

5.1. Resulting Network Metrics.

5.1.1. We have proposed network metrics which are generalized entropy functions of the Markov monoid matrix M generated by an altered connection matrix C. When sorted, the associated entropy spectra for the columns and rows of C monitor the state and time evolution of the incoming and outgoing entropy at network nodes.

5.1.2. These well defined functions, along with the network amplitude function, satisfy our original criteria for network metrics. They can be used to dynamically monitor networks relative to such normal metrical values thus identifying when the network statistically alters its intrinsic patterns of connectivity.

5.2. Further Analysis Needed.

5.2.1. Although these functions provide an extremely promising set of metrics for the characterization of networks in general, we have only begun to explore the association of specific spectral form and behavior with specific topologies and network behaviors.

5.2.2. The spectral form of the incoming and outgoing entropy functions is anticipated to be very different in different applications and the study of such spectral properties will be extensions of the scale-free network structures that have been recently investigated 14, 15, 19]. This is our first major new need for further research: It will involve some purely mathematical analysis and some testing on diverse networks to determine their normal entropy spectral forms. This experimental research need not be done in real time but can be accomplished off line.

5.3. Difficulties Encountered with Real Time Deployment.

5.3.1. Additionally, we have found great difficulty in running the analysis on networks long after the fact using transmitted data that is stored for later analysis. The major difficulty in this arises because the system is not identifying the anomaly in real time so that it cannot be easily correlated with other network system tracking software and observed network parameters.

5.3.2. We have not yet had time (after the completion of the code and its testing) to successfully negotiate the placement of our software into meaningful dynamical environments in real time due to political and privacy concerns. This is our second major need for future research and it must be done on line and in real time in order to optimally correlate anomalies with other indicators and information.

6. References:

1. Herbert Goldstein. *Classical Mechanics* Addison-Wesley 2005 This classic reference provides a full rigorous background on the laws classical mechanics.

2. Albert Messiah. *Quantum Mechanics Vol 1 & 2* John Wiley & Sons 1961. Likewise this two volume set provides both the physical and mathematical backgrounds for quantum theory.

3. Kerson Huang. *Statistical Mechanics* John Wiley & Sons 1963. Both thermodynamics and statistical mechanics foundations are covered in this classic test.

4. Morton Hamermesh. Group Theory Addison-Wesley 2005. This text provides a full background on group theory of both discrete and continous groups.

5. Joseph E. Johnson. Networks, Markov Lie Monoids, and Generalized Entropy, p129 in
6. Vladimir Gorodetsky, Igor Kotenko, Victor Skormin (Eds.). Computer Network Security: Third International Workshop on Mathematical Methods, Models, and Architectures for Computer Network Security, MMM-ACNS 2005, St. Petersburg, Russia, September 2005 Proceedings

7. Joseph E. Johnson. *Markov-type Lie groups in GL(n,R),* J. Math. Phys. 26 (2) 1985. This was the foundational original work that led to the linkages in this current investigation among networks, Markov transformations, and continuous group and algebra theory. The results are summarized in the technical sessions.

8. Alagna, Chen, et al (Eds) *The Black Book on Corporate Security* Larstan Publishing Washington DC (2005). This work gives an overview of the diverse threats via internet attacks.

9. O. Tarakanov, V.A. Skormin, & S.P.Sokolova. *Immunocomputing Principles And Applications,* Springer NY (2003)

10. Renyi. *Probability Theory*, North-Holland Series in Applied Mathematics and Mechanics, North-Holland Publishing Co (1070)

11. Renyi. *Selected Papers of Alfred Renyi,* Akademia Kiado, Buadapest, Vol 2 of 3 volumes (1976)

12. Vladimir Gudkov & Joseph E. Johnson. *Networks as a Complex System: Information Flow Analysis,* arXiv:lin.CD/0110008vl (2001)

13. Vladimir Gudkov & Joseph E. Johnson. *Multidimensional Network Monitoring for Intrusion Detection*, arXiv:cs.CR/020620vl (2002)

14. Reka Albert and Albert-Laszio Barabasi. *Statistical Mechanics of Complex Networks*, Reviews of Modern Physics, Vol 74, 47, (2002)

15. Albert-Laszlo Barabasi and Eric Bonabeau. *Scale Free Networks*, Scientific American, 50 May (2003)

16. Patrick Radden Keefe. Can Network Theory Thwart Terrorists?, NY Times Magazine Section, p16, March 12 (2006)

17. M Ayedemir et al. Computer Networks 36 (2001)

18. S. Northcutt, J. Novak, and D. Mc Lachlan. Network Intrusion Detection, An Analyst's Handbook, New Riders Publishing, Indiapolis IN (2001)

19. Lun Li, David Alderson, John C Doyle, Walter Willinger. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications. August 25, (2005) preprint. This paper contains over 100 references providing an extensive current background on graph theory.

20. Yaneer Bar-Yam, New England Complex Systems Institute, International Conference on Complex Systems – Boston MA, June 25-30 2006 Proceedings. These proceedings were not yet published at the time of this report but the abstracts of papers provide evidence of the vast spectrum of interdisciplinary interest in network and graph theory in science, mathematics, and complex systems.